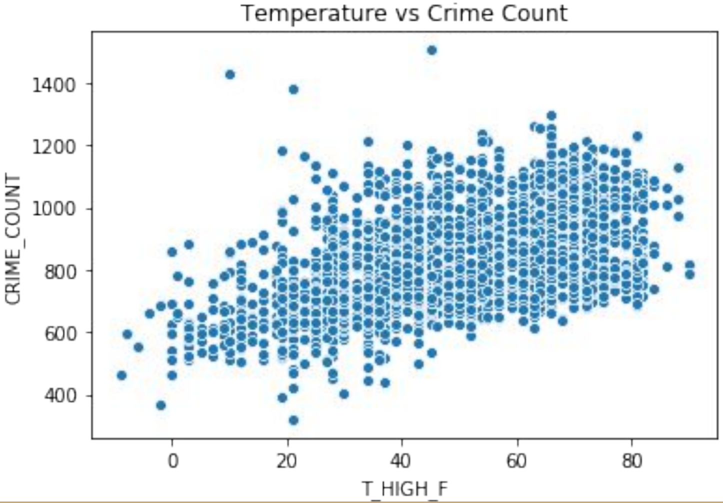
Chicago Daily Crime Count Prediction Using Regression Model

Project 2 : Luther

Matthew Lee
7/18/2019

Foundation of the Project

- Hypothesis
  - Weather has correlation to crime rates.
    - ex) When temperature is really high, people get hot-headed and impatient easily.
- Objective
  - Develop a regression model to predict total daily crime counts reliably.



What's the Business Value of this Model?

- The City of Chicago spends more than $4 million dollars on the Chicago Police Department **DAILY.**

- Assist Chicago city with better budgeting.

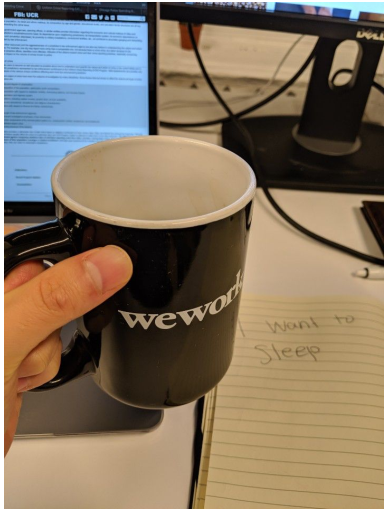
What Are the Ingredients for the Model?

Recommended Recipe by FBI

- Population density and degree of urbanization.
- Variations in composition of the population, particularly youth concentration.
- Stability of the population with respect to residents' mobility, commuting patterns, and transient factors.
- Modes of transportation and highway system.
- Economic conditions, including median income, poverty level, and job availability.
- Cultural factors and educational, recreational, and religious characteristics.
- Family conditions with respect to divorce and family cohesiveness.
- Climate.
- Effective strength of law enforcement agencies.
- Administrative and investigative emphases of law enforcement.
- Policies of other components of the criminal justice system (i.e., prosecutorial, judicial, correctional, and probational).
- Citizens' attitudes toward crime.
- Crime reporting practices of the citizenry.

What Are the Ingredients for the Model?

Only-Got-2-Weeks Recipe

- Weather Data
- Bus & Rail Boarding Ridership Count
- Unemployment Rate Data
- 

Tools Used to Cook This Model

- Data
  - 2798 Rows (2009~2018)
- Features
  - Numerical : 7
  - Categorical : 3
  - len(get_dummies(Categorical)) : 24
- Preprocessing Tools
  - PolynomialFeatures
  - StandardScaler
- Model & Model Selection
  - Lasso (Baseline Modeling)
  - ElasticNet (Lasso & Ridge)
  - RandomizedSearchCV (Hyperparameter Tuning)

Feature Engineering

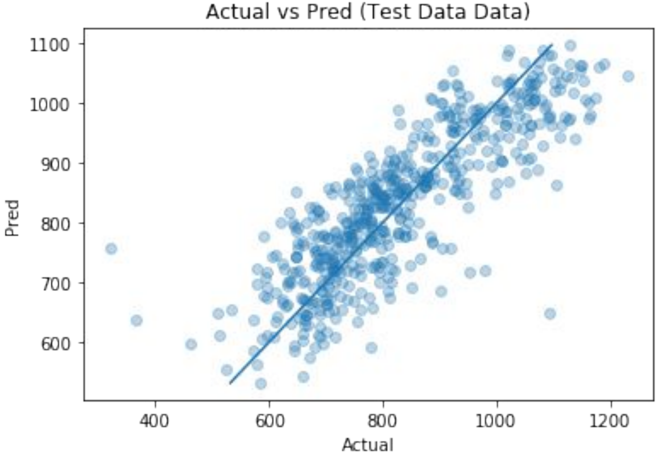
- Moving average of temperature
- Day name of the week : Mon ~ Sun
- Name of month : Jan ~ Dec
- Weather Description : Clear, Cloudy, Rainy, ...

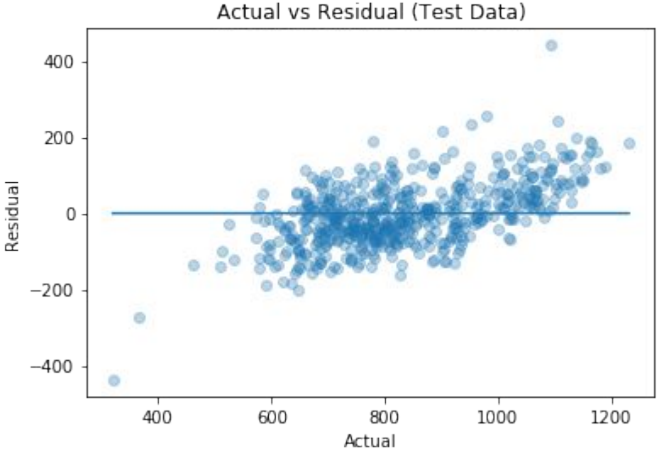
Quality Inspection & Control of the Model

- Pairplots, heatmaps, Prediction vs Actual, and Residual plots
- R^2 scoring for validating pre-processing approaches, feature selection, and model selection processes.

Results & Insights - Base Lasso Model

- Baseline Lasso model with alpha=1. Just numerical features.





	Weather		Weather Ridership		Weather Ridership Unemployment		Unemployment	
	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
R^2_Train	0.23	0.28	0.25	0.35	0.67	0.84	0.43	0.40
R^2_Test	0.28	0.30	0.33	0.41	0.70	0.82	0.40	0.44

Results & Insights - Base Lasso Model Cont.

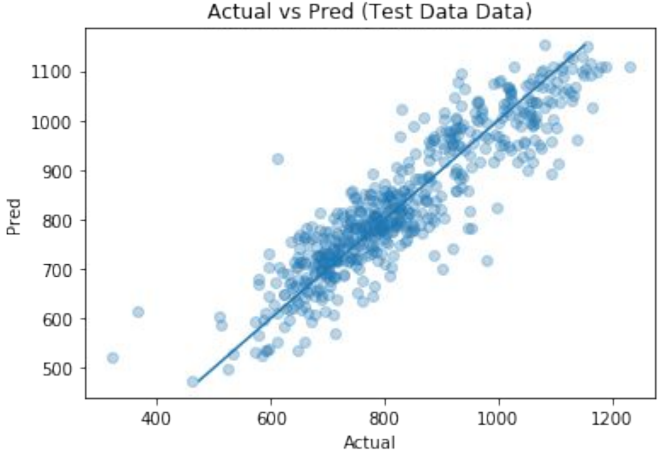
- Lasso model's coefficients showed that there were direct relationships between temperature, total ridership, unemployment and daily crime rates in Chicago.
- Humidity and wind showed inverse relationship with daily crime counts in Chicago.

*Actual coefficients in appendix

	Weather		Weather Ridership		Weather Ridership Unemployment		Unemployment	
	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
R^2_Train	0.23	0.28	0.25	0.35	0.67	0.84	0.43	0.40
R^2_Test	0.28	0.30	0.33	0.41	0.70	0.82	0.40	0.44

Results & Insights - ElasticNet

- ElasticNet with RandomizedSearchCV



	Weather		Weather Ridership		Weather Ridership Unemployment		Unemployment	
	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
R^2_Train	0.23	0.28	0.25	0.35	0.67	0.84	0.43	0.40
R^2_Test	0.28	0.30	0.33	0.41	0.70	0.82	0.40	0.44

Results & Insights - ElasticNet Cont.

- Coefficients from ElasticNet picked up temperature, total ridership, and unemployment as key features affecting daily crime counts in Chicago as well.
    - There were total 595 features after applying polynomial features.

*Actual coefficients in appendix

	Weather		Weather Ridership		Weather Ridership Unemployment		Unemployment	
	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
R^2_Train	0.23	0.28	0.25	0.35	0.67	0.84	0.43	0.40
R^2_Test	0.28	0.30	0.33	0.41	0.70	0.82	0.40	0.44

Conclusion & Further Expansion

- Both baseline Lasso model and ElasticNet were successful picking up meaningful signal from data.
- Even though ElasticNet scored R^2 value, I would recommend baseline Lasso model with key 7 features due to its strong interpretability.
    - MAE_base = 64.45 & MAE_ElasticNet = 50.08
- Model should be tested in other city’s data to test how well it generalizes.
- More data could be incorporated such as demographics data.

thank you

Appendix

- All Features
- Hyperparameters for PolynomialFeatures & ElasticNet & RandomizedSearchCV
- Baseline Model & Final Model Coefficients
- Pairplot of Categorical Features
- Heatap of Categorical Features
- Outliers

All Features

- **Base Features (Numerical) :** 'T_HIGH_F', 'T_LOW_F', 'HUMIDITY_%', 'BAROMETER_HG', 'WIND_MPH', ‘TOTAL_RIDES', 'UNEMP_RATE'

- **Engineered Feature (Numerical) :** 'T_HIGH_F_MOVING_AVG'

- **Engineered Features (Categorical) :** 'Friday', 'Monday', 'Saturday', 'Sunday', 'Thursday', 'Tuesday', 'Wednesday', 'April', 'August', 'December', 'February', 'January', 'July', 'June', 'March', 'May', 'November', 'October', 'September', 'Clear.', 'Fog.', 'Mostly cloudy.', 'Other', 'Overcast.', 'Passing clouds.'

Hyperparameters

- PolynomialFeatures()
  - Degree : [1,2]
- ElasticNet()
  - alpha : 10**np.linspace(-2,2,300)
  - l1_ratio : np.linspace(0,1,100)
- RandomizedSearchCV()
  - cv = 5
  - scoring = r2
  - Random_state = 209

Hyperparameters Cont.

Final hyperparamters chosen by RandomizedSearchCV

```
RandomizedSearchCV

R^2 Score (Train Data):  0.8445360055545827
R^2 Score (Test Data):  0.8169573186475456

Best params :  {'polynomial__degree': 2, 'model__random_state': 209, 'model__max_iter': 2000, 'model__l1_ratio': 0.94
94949494949496, 'model__alpha': 0.04665479882234473}
```

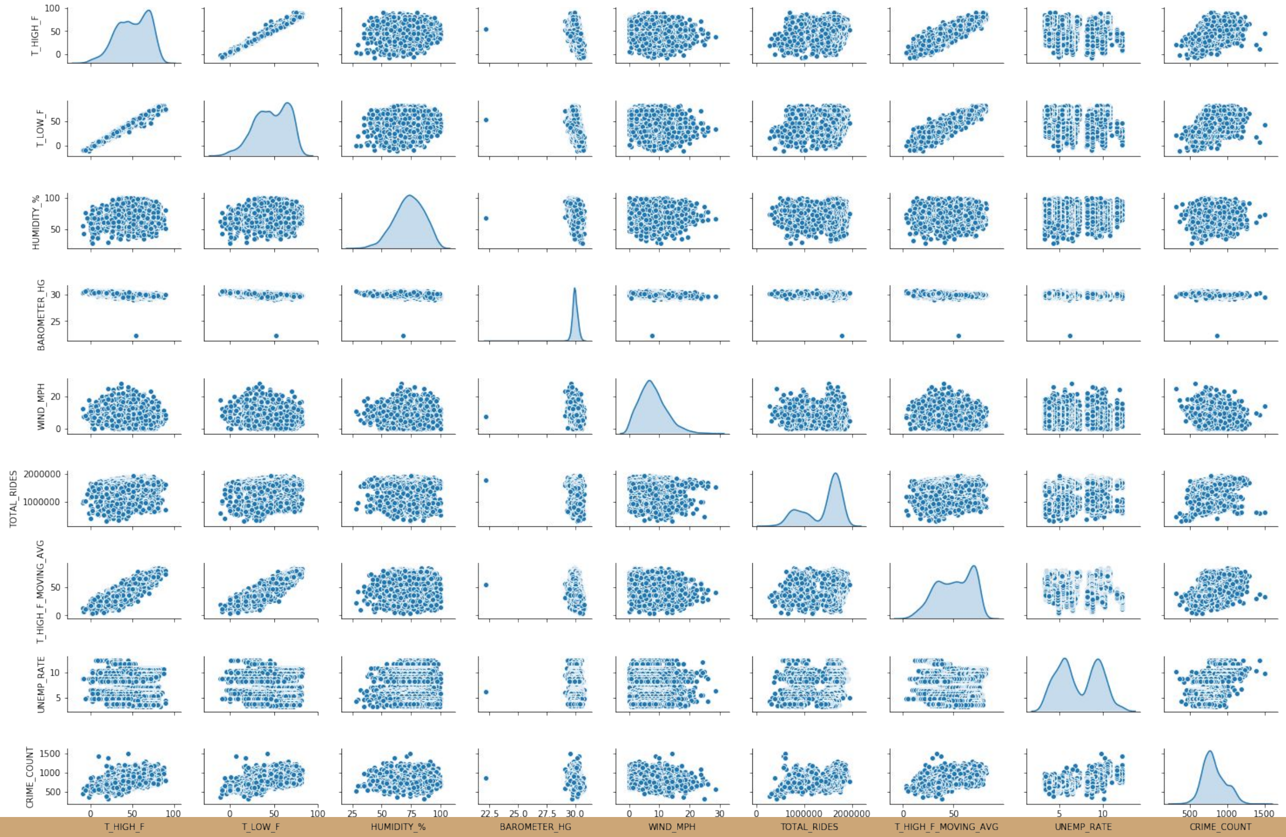
Final Coefficients - Lasso

```
Base Model

Base_R^2_Train :  0.6733373787242718
Base_R^2_Test :  0.7027140682440205

Coefficients :
('T_HIGH_F', 45.82062524313216)
('T_LOW_F', 19.526820913939)
('HUMIDITY_%', -3.707311114552946)
('BAROMETER_HG', 0.0)
('WIND_MPH', -10.838539034006782)
('TOTAL_RIDES', 16.12684061071973)
('UNEMP_RATE', 96.55139115496372)
```

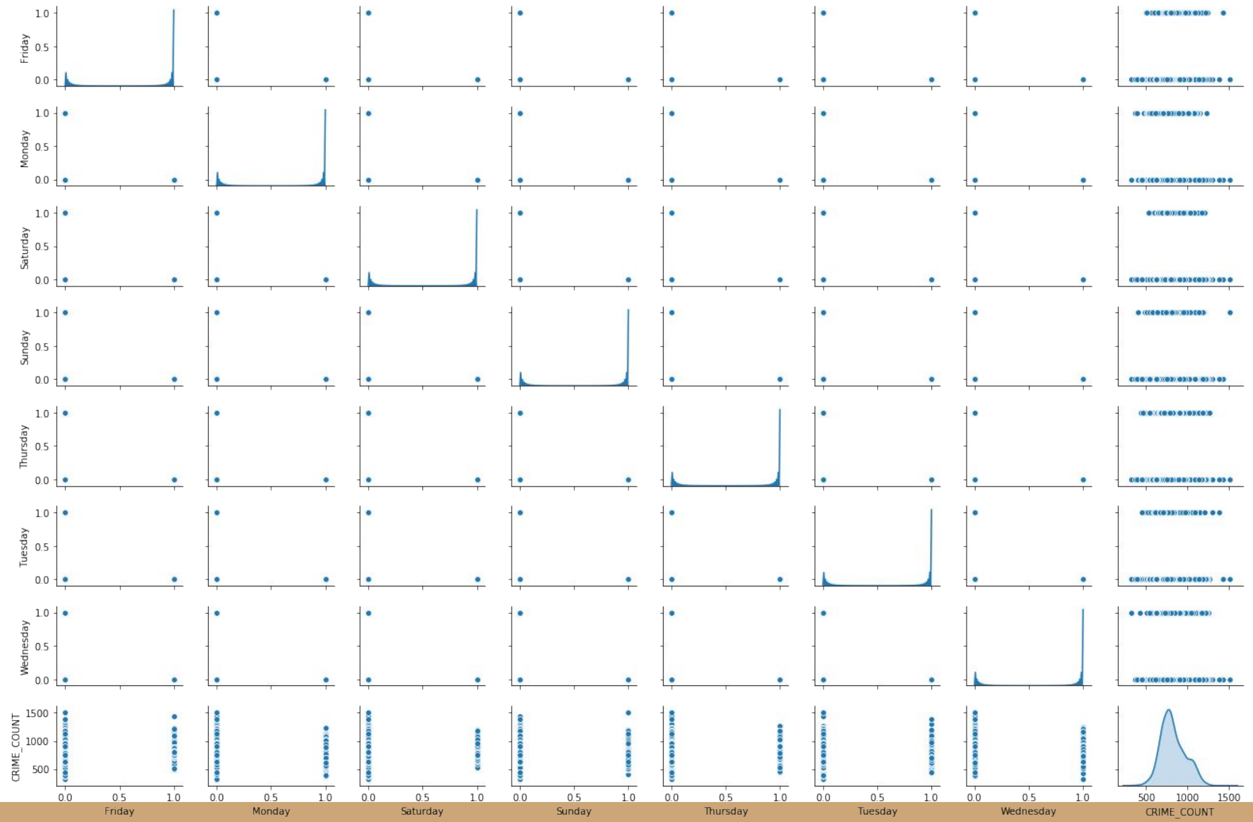
Significant Coefficients - ElasticNet

	COEF	COEF_VALUE
217	UNEMP_RATE^2	236.197493
201	TOTAL_RIDES December	34.984679
194	TOTAL_RIDES Saturday	30.790255
190	TOTAL_RIDES UNEMP_RATE	27.355908
189	TOTAL_RIDES^2	23.588516
195	TOTAL_RIDES Sunday	22.852465
258	T_HIGH_F_MOVING_AVG June	22.379125
60	T_HIGH_F September	20.072442
204	TOTAL_RIDES July	-21.717573
229	UNEMP_RATE February	-23.261130
221	UNEMP_RATE Saturday	-23.288782
210	TOTAL_RIDES September	-29.127306
103	HUMIDITY_% UNEMP_RATE	-34.092502
7	UNEMP_RATE	-65.881480
133	BAROMETER_HG UNEMP_RATE	-66.425533
203	TOTAL_RIDES January	-99.225158

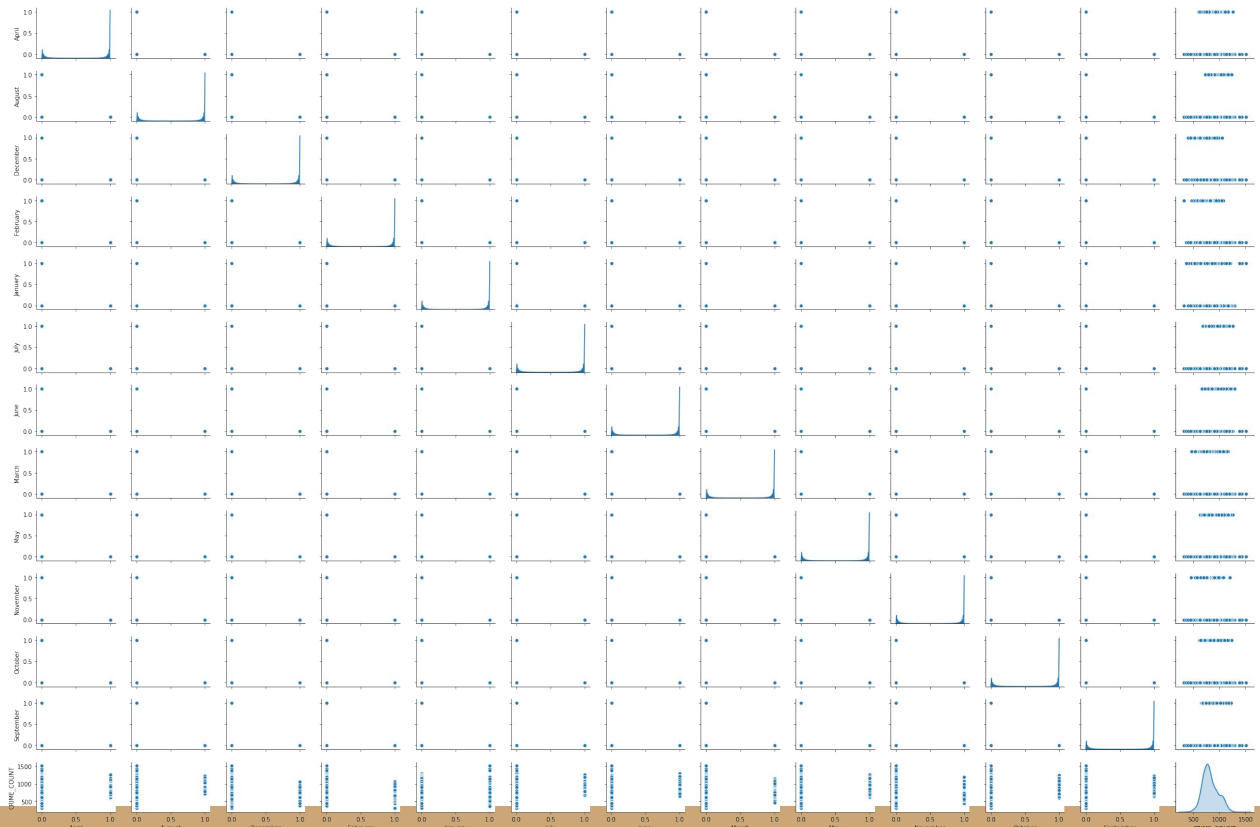
Pair Plot : Numerical Features



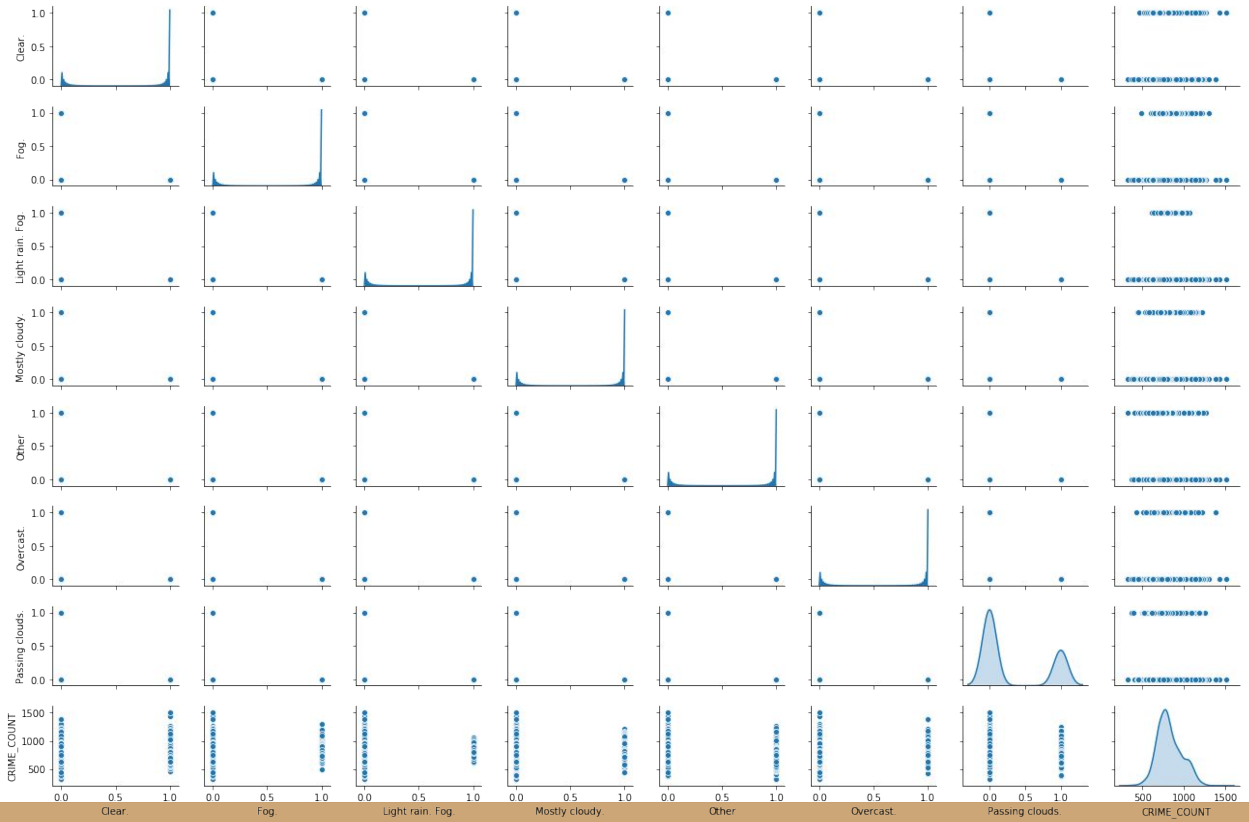
Pair Plot : Categorical Features - Days



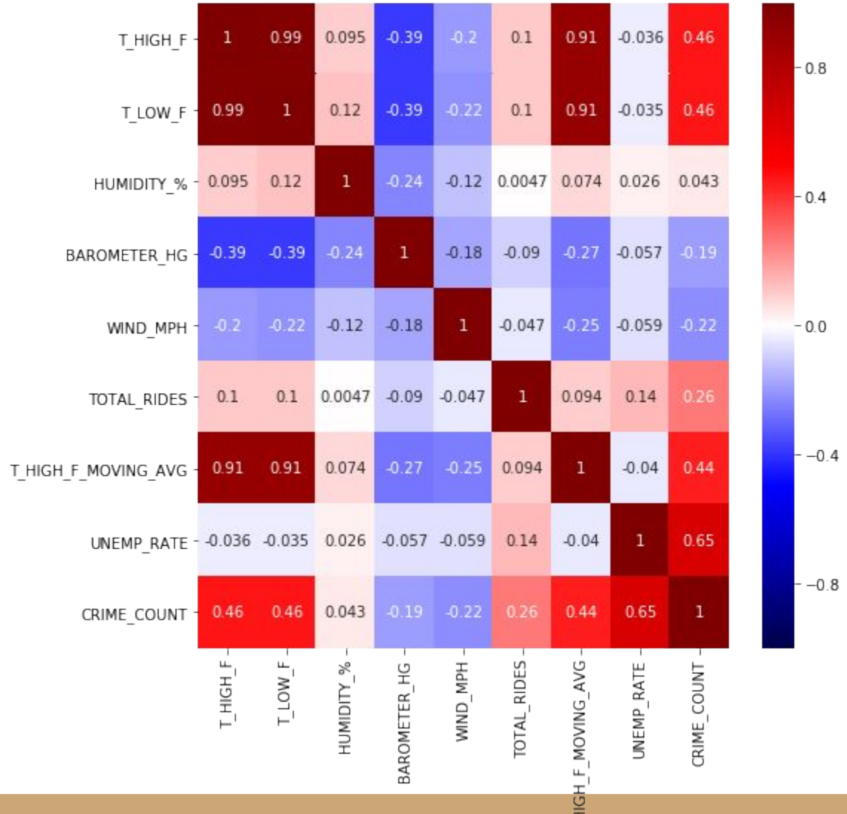
Pair Plot : Categorical Features - Months



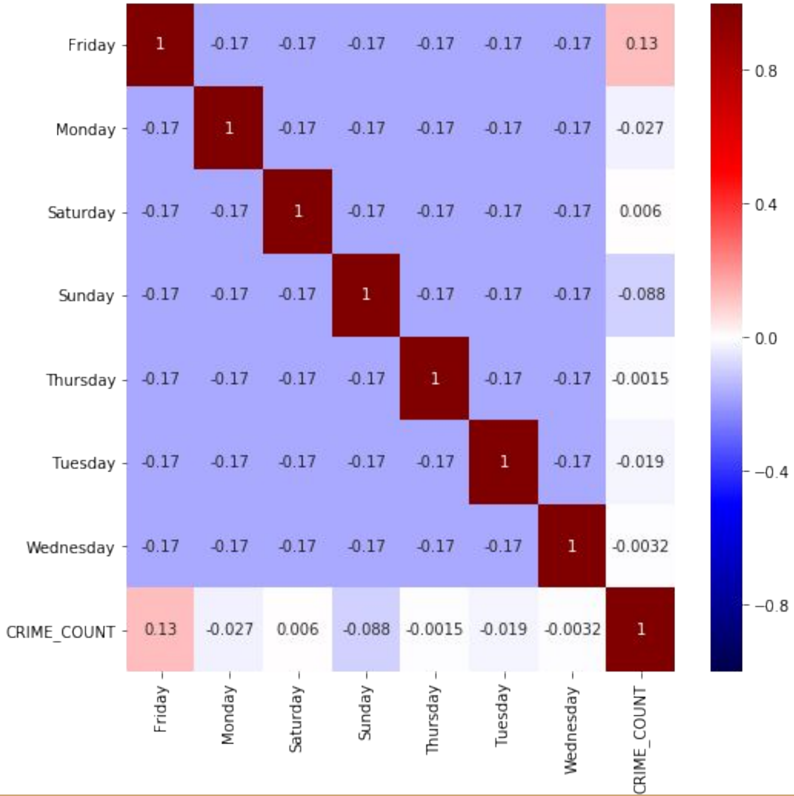
Pair Plot : Categorical Features - Weather Desc.



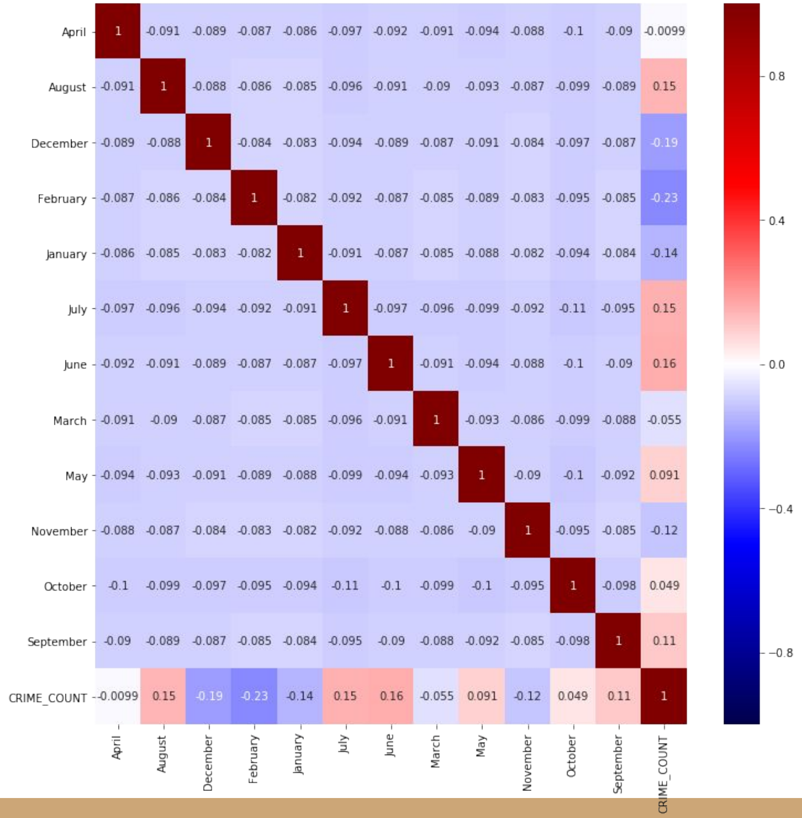
Heatmap - Numerical Features



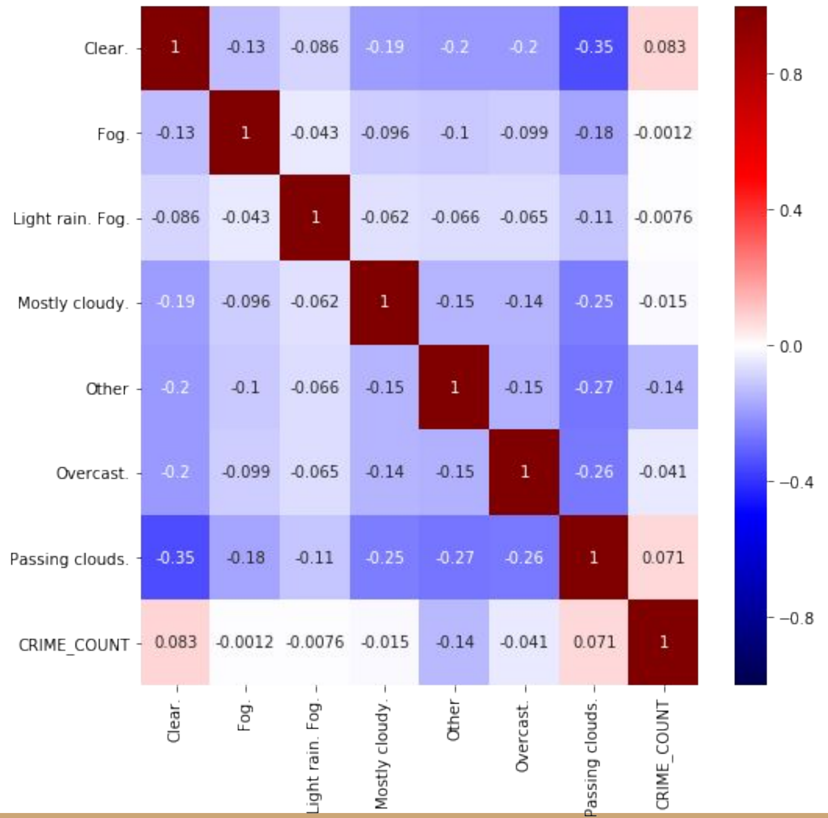
Heatmap - Categorical Features - Days



Heatmap - Categorical Features -Months



Heatmap - Categorical Features -Months



Outliers

ROCESSED	T_HIGH_F	T_LOW_F	HUMIDITY_%	BAROMETER_HG	WIND_MPH	TOTAL_RIDES	UNEMP_RATE	T_HIGH_F_MOVING_AVG	RESIDUAL	CRIME_COUNT
2018-11-01	52.0	52.0	58.0	29.90	9.321	1664628.0	3.5	45.857143	201.281247	902.0
2018-05-01	64.0	57.0	35.0	29.97	9.943	1654457.0	3.4	50.857143	261.667259	978.0
2014-01-06	-2.0	-11.0	63.0	30.16	18.021	580617.0	8.7	19.285714	-248.984205	366.0
2016-01-01	25.0	19.0	67.0	30.25	15.535	623156.0	6.5	42.000000	200.504367	1092.0
2011-02-02	21.0	21.0	86.0	29.75	24.857	428521.0	10.1	38.571429	-202.848361	320.0
2017-01-02	36.0	32.0	74.0	30.15	6.836	686663.0	5.8	38.571429	-311.115876	612.0

	T_HIGH_F	T_LOW_F	HUMIDITY_%	BAROMETER_HG	WIND_MPH	TOTAL_RIDES	UNEMP_RATE	T_HIGH_F_MOVING_AVG	RESIDUAL	CRIME_COUNT
count	560.000000	560.000000	560.000000	560.000000	560.000000	5.600000e+02	560.000000	560.000000	560.000000	560.000000
mean	52.537500	49.105357	72.816071	30.011750	7.957407	1.415746e+06	7.349821	52.401020	-1.232339	830.869643
std	20.152159	19.784202	11.639490	0.219882	4.500470	3.811230e+05	2.346146	18.608802	65.564835	153.275107
min	-8.000000	-11.000000	35.000000	29.240000	0.000000	4.285210e+05	3.400000	3.142857	-311.115876	320.000000
25%	37.000000	34.000000	65.000000	29.880000	4.971000	1.075809e+06	5.400000	38.000000	-41.504705	713.750000
50%	54.000000	52.000000	73.000000	29.990000	7.457000	1.586935e+06	7.100000	54.785714	-4.839720	807.000000
75%	70.000000	66.000000	81.000000	30.152500	11.185000	1.696251e+06	9.500000	69.428571	37.258565	935.000000
max	88.000000	82.000000	98.000000	30.810000	28.585000	1.926454e+06	12.200000	82.142857	261.667259	1231.000000

Feedback from Instructors

- At the end, tie back to the original business value.
- Change format of data in table to easily digestible format. Maybe use flow diagram/other visuals
- Tie MAE score back to the original target values to provide some context about magnitude/impact of MAE scores. Help audience build the intuition.